3-D Atomistic Nanoelectronic Modeling on High Performance Clusters: Multimillion Atom Simulations

GERHARD KLIMECK*, FABIANO OYAFUSO†, R. CHRIS BOWEN*,
TIMOTHY B. BOYKIN†, THOMAS A. CWIK*, EDITH HUANG*, and
EDWARD S. VINYARD*,

* *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109*
† *University of Alabama in Huntsville, AL*

Electronic device scaling is ultimately limited by atomic dimensions. The simulation of electronic structure and electron transport on these length scales must be fundamentally quantum mechanical. This leads to computational models that account for fundamental physical interactions using an atomistic basis and tax even the largest available supercomputer when simulating measurable devices. The prototype development of a software tool that enables this class of simulation is presented. Realistically sized structures contain one million to tens of millions atoms that need to be represented with an appropriate basis. The resulting sparse complex Hamiltonian matrix is of the order of tens of millions. A custom matrix vector multiply algorithm that is coupled to a Lanczos and/or Rayleigh-Ritz eigenvalue solver has been developed and ported to a Beowulf cluster as well as an Origin 2000. First benchmarking results of these algorithms as well as the first results of quantum dot simulations are reported.

Keywords: quantum dot, nanoelectronics, sparse matrix-vector multiplication

1. Introduction

The goal of reducing payload in future space missions while increasing mission capability demands miniaturization of measurement, analytical, and communication systems. The ultimate scaling limit of individual semiconductor devices are atomic dimensions. The enabling technology for miniaturization of space mission electronics has been the miniaturization of semiconductor devices (Fig. 1a adopted from the SIA Roadmap[1]) of the past 40 years. The development has surpassed every expectation and overcome (so far) every predicted technological obstacle. It has become evident that not technology but the *atomic dimensions* of the underlying crystalline lattice and the countable *electron number* (Fig. 1b) ultimately limit[2] this scaling trend. A variety of novel nano-scaled detector[3] and computation[4] schemes based on quantum dots have been proposed and/or demonstrated. The work presented here is aimed at providing a simulation tool that enables the fundamental modeling of electron devices on the length scales of a few nanometers. The problem size and the choice
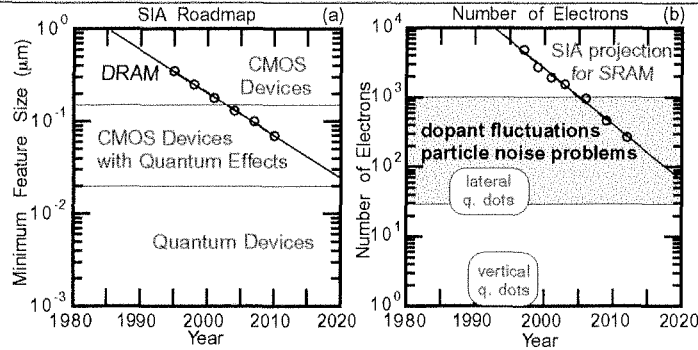
Figure 1. (a) Minimum 2D feature size as projected on the SIA roadmap [1]. Layer thickness od $0.01\mu m$ in the next generation devices are not captured in this graph. (b) Number of electrons under a CMOS SRAM gate[2]. Dopant fluctuation and particle noise fluctuations may make reliable circuit design impossible, since each device may vary from the next significantly.

of atomistic basis sets results in numerical representations that require the usage of supercomputers[5].

Simulation for device design and characterization: Physics-based device simulation has penetrated the mainstream semiconductor industry device design and characterization process[6] in the last one or two years. This penetration can be attributed to the increased experimental costs of nano-scale layer and feature characterizations and the introduction of new materials. The modeling - fabrication - characterization triangle that has existed in circuit design for the past decade has now established itself for the underlying semiconductor device design as well. This coupled process will enable the device design for the next device generations if the physics-based simulation tools can deliver the needed accuracy. As the electronic device sizes shrink further towards the tens of nanometer size scale (deca-nano) in all three dimensions, new physical phenomena will emerge, that previously could be safely ignored. These phenomena are based on the quantum mechanical nature of electrons which is typically ignored by or patched into existing commercial device simulators. Such an approach will probably suffice for the next two device generations with empirical model calibration and limited scaling projection capability. In order to correctly model physically observed effects such as electron tunneling, state quantization and charge quantization more sophisticated models need to be developed and incorporated into device simulators that can handle realistically sized systems.

Our modeling agenda: Following out 1-D Nanoelectronic modeling work[7, 8] (NEMO) we are developing an atomistic-based, nanoelectronic modeling tool (NEMO-3D). The following section discusses some of the system size issues and quantum mechanical modeling capabilities that need to be included into such a simulator.

2. Quantum Dot Applications and Modeling Requirements

Quantum devices are not shown on the SIA roadmap for lithography because it focuses on pure silicon device scaling. In particular Figure 1a only shows the lateral feature size, while layer thicknesses are already on the order of $0.01\mu m$. Various quantum dot implementations in different material systems as well as silicon have been examined since the late 1980's (Fig. 1b), and several designs have shown room temperature operation. Pyramidal self-assembled quantum dot arrays in particular are promising candidates to be used in quantum well lasers and detectors[3] within a few years.

What is a Quantum Dot? A quantum dot (QD) can be described as a solid state structure in which a (small) number of electrons are isolated from the surrounding environment. This is achieved by "placing" an electrical insulator around a (semi-)conductor. If the central region is small and clean enough effects due to state and charge quantization can be measured macroscopically. QDs can therefore be viewed as *artificial atoms*. They represent the ultimate limit of scaled solid state devices. The large parameter space of material systems, shapes, and doping profiles allows for a detailed engineering of the electrical and optical properties of quantum dots. In particular, the fine tuning of optical transition energies applies to JPL's immediate interest in far infrared detectors and emitters.

Near term quantum dot applications: Currently several research groups are incorporating self-assembled InAs QDs into AlGaAs quantum wells in optical detector and laser structures[3]. The reduced degree of freedom reduces the scattering of the confined electrons and therefore increases the state lifetimes. This increase promises better device performances such as reduced threshold currents, decreased linewidths, reduced dark currents[9] and higher operating temperatures. These are systems which apply to JPL's immediate interest in far infrared detectors and emitters.

Size and atom number estimates of realistic systems: The modeling of an individual self-assembled InAs quantum dot of 30nm diameter and 5nm height embedded in GaAs of buffer width 5nm requires roughly a simulation domain of $40 \times 40 \times 15 nm^3$ containing approximately 1 million atoms. A horizontal array of four such dots separated by 20nm requires roughly a simulation domain of $90 \times 90 \times 15 nm^3$ which corresponds to 5.2 million atoms. A $70 \times 70 \times 70 nm^3$ cube of Silicon which might be needed to simulate ultra-scaled CMOS device contains about 15 million atoms. There are about 43 atoms in $1 nm^3$ as a rule of thumb.

3. Basis Representation

Researchers have explored a variety of different approaches to represent matter in a nano-scaled system. All these approaches fall into two major categories: atomistic and non-atomistic. The non-atomistic approaches do not attempt to model each individual atom in the structure, but introduce a variety of different approximations that are usually based on a continuous, jellium-type description of matter. Such approaches typically deal on the lowest level with effective masses and band edges. The popular $k \cdot p$ approach belongs in this category. These approaches do not contain any crystalline information and are fundamentally not well suited for the atomistic representation of

nano-scale features such as interfaces and disorder. Atomistic approaches attempt to include the electronic wavefunction of each atom in some approximation. The critical question is what basis set to use for the representation of the electronic wavefunction. There are two schools of thought: 1) plane waves[5], and 2) local orbitals[10]. A list of pros and cons for each method could be presented; instead we just state here that we consider both methods complementary to each other and that we choose the local orbital basis (tight-binding) approximation for its ability to model finite structures (not infinitely extended) and for its past success[7, 8] in quantum mechanical modeling of electron transport.

The basic idea of the tight-binding method is that one selects a basis consisting of atomic orbitals (such as s, p, and d) to create a single electron Hamiltonian that represents the bulk electronic properties of the material. The interactions between the different orbitals within an atom and between nearest neighbor atoms are treated as empirical fitting parameters. A variety of parameterizations of nearest neighbor and second-nearest neighbor tight binding models have been published including different orbital configurations[11–16]. Our simulator typically uses an sp3s* or sp3s*d5 model that consists of 5 or 10 spin degenerate basis states, respectively. Each atom is therefore represented by a 10×10 or 20×20 matrix. We have limited ourselves to nearest neighbor interactions to enable a simpler interaction representation in the presence of strain. This nearest neighbor model restricts the number of non-zeros per row to a small value independent of device size and yields a sparse, block banded Hamiltonian with $O(n)$ non-zeros, where n is the number of atoms. The strain that arises near interfaces of materials with different lattice constants yields a different, position-dependent coupling between each neighboring pairs of atoms. For the case of a zinc-blende lattice, each atom has four neighbors. Therefore, the storage requirement for 1 million atoms can be estimated as 10^6 *atoms* $\times 5$ *diagonals* $\times (20 \times 20$ *basis*) $\times 16 bytes/2 (for Hermiticity) = 16GB$. Algorithms that use more of the symmetry of the Hamiltonian are considered for future development. Currently we have the option to store the Hamiltonian or to recompute it on the fly.

Quantum dots are characterized by confinement in all three spatial dimensions, so that the Hamiltonian no longer commutes with *any* of the (discrete) translation operators. The wavevector is hence *not* a conserved quantity in *any* dimension. The most appropriate basis for representing such a highly confined wavefuntion is, therefore, one consisting of atomic-like orbitals centered on each atom of the crystal. Following Slater and Koster [18], we take the atomic-like orbitals to be orthonormal. We consider a crystal whose Bravais lattice points are given by

$$\mathbf{R}_{n1n2n3} = n_1\mathbf{a}_1 + n_1\mathbf{a}_2 + n_1\mathbf{a}_3 \tag{1}$$

where the $\mathbf{a}_i$ are primitive direct-lattice translation vectors and the n_i are integers. If there is more than one atom per cell (as is the case with, for example, GaAs or Si) we index the atoms within a cell by μ and the location of the μth atom within the cell located at 1 is given by $\mathbf{R}_{n1n2n3} + \mathbf{v}_\mu$ where $\mathbf{v}_\mu$ is the displacement relative to the cell origin. We normalize the wavefunction over a volume consisting of N_i cells in the $\mathbf{a}_i$ $(i = 1, 2, 3)$ direction and write the state a general expansion in terms of localized atomic-like orbitals:

$$|\Psi> = \frac{1}{\sqrt{N_1 N_2 N_3}} \sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} \sum_{n_3=1}^{N_3} \sum_\alpha \sum_\mu C_{n_1 n_2 n_3}^{(\alpha\mu)} |\alpha; \mu; \mathbf{R}_{n1n2n3} + \mathbf{v}_\mu> \tag{2}$$
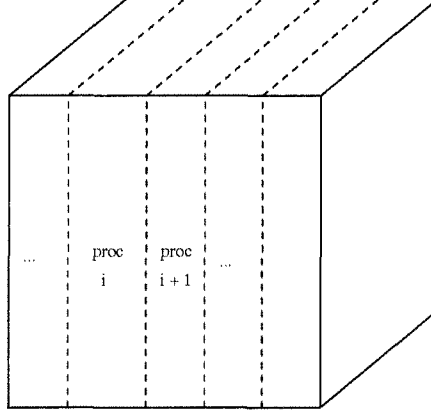
Figure 2. The device is decomposed into slabs (layers of atoms) which are directly mapped to individual processors.

In Eq.(2) α indexes the atomic-like orbitals centered on the μ atoms within each cell $(n_1 n_2 n_3)$. The Schrödinger equation thus appears as a system of simultaneous equations given by:

$$<\alpha; \mu; \mathbf{R}_{n1n2n3} + \mathbf{v}_\mu|(\hat{\mathcal{H}} - E\hat{1})|\Psi> = 0 \tag{3}$$

In Eq.(3) we express the matrix elements between localized orbitals as tight-binding parameters, in our case limiting the interactions to nearest-neighbor.

4. Numerical details

4.1. Parallel implementation of sparse matrix-vector product

The goal of the simulation is to solve the eigenvalue problem for low lying electron and hole states near the bandedge. Two algorithms, a Raleigh-Ritz minimization algorithm and a Lanczos method, have been parallelized to solve this problem. At the heart of each is a sparse matrix-vector multiplication. For implementation on a distributed memory platform, data must be partitioned across processors so as to facilitate this fundamental operation. For good load balance the device is partitioned into approximately equally sized sets of atoms which are mapped to individual processors. Because only nearest neighbor interactions are modeled, a naive partition of the device by parallel slices creates a mapping such that any atom must communicate with neighbors that are at most one processor away. This scheme, shown in Figure 2, lends itself to a 1D chain network topology and results in a block-tridiagonal Hamiltonian, where each block corresponds to a pair of processors, and each processor holds the column of blocks associated with its atoms. Communication costs, roughly proportional to the boundary separating these sets, scales only with surface area $(O(n^{2/3}))$ rather than with volume $(O(n))$, where n is the number of atoms. In a matrix-vector multiply, both the sparse Hamiltonian and the dense vector are partitioned among processors in an intuitive way; each processor p holds unique copies of both the nonzero matrix
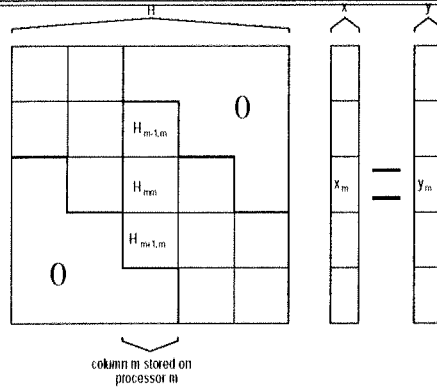
Figure 3. Example matrix vector multiplication on 5 processors performed in a columnwise fashion, so that the mth block column and section x_m are stored on processor m. The nearest neighbor model with non-periodic boundary conditions guarantees that the Hamiltonian is block-tridiagonal, so that communication is performed only with nearest neighbor processors.

elements of the sparse Hamiltonian associated with the orbitals of the atoms mapped to processor p and also the components of the dense vector associated with atomic orbitals mapped to p. The matrix-vector multiply is performed in a columnwise fashion as shown in Fig. 3. That is, processor j computes

$$y_{i,j} = H_{i,j} x_j \qquad (i = j, j \pm 1) \qquad (4)$$

where $H_{i,j}$ is the block of the Hamiltonian associated with nodes i and j, and x_j are the components of x stored locally on node j. The result of the matrix-vector multiplication of the off-diagonal blocks $(i \neq j)$ with the local portion of the dense vector is sent via MPI calls to neighboring processors in two steps. First, all but the rightmost processor send data to the right (and receives data from the right, possibly in full duplexing depending on the actual MPI implementation), subsequently followed by communication to the left neighbors. In addition, not all rows of the off-diagonal blocks are in general non-zero, although the sparse structure of these blocks depends on the particular crystal structure in question. In practice, however, a sufficient fraction of zero rows exist that compressing the matrix-vector multiplication by removing structurally guaranteed zeros is worthwhile despite the additional level of indirection required to keep track of the non-zero structure.

4.2. Performance

In this section, we discuss the performance of NEMO 3-D using the parallelized Lanczos algorithm on two platforms, a Beowulf commodity cluster of Pentium III's and a shared memory SGI Origin 2000. Figure 4a displays the performance of 30 iterations of the Lanczos solver on a Beowulf system consisting of 32 nodes with two 933 MHz Pentium III CPUs and 1 GB of RAM per node. Different processors communicate using a non-shared memory MPI implementation and over a slow 100 base-T ethernet connection. Curves corresponding to five different problem sizes are shown. Dashed curves indicating a linear (ideal) scaling are also shown for reference.
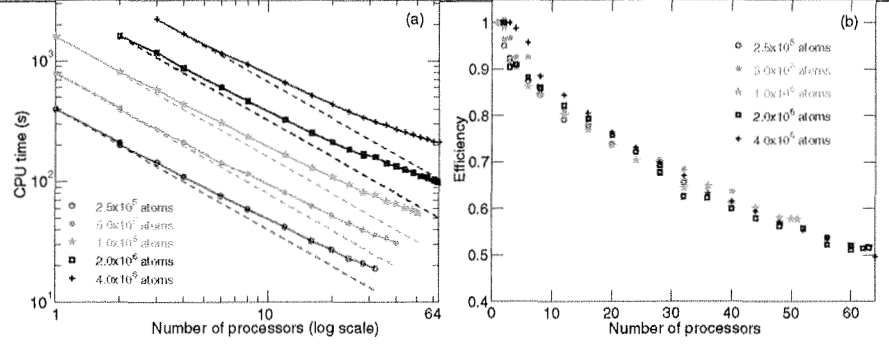
Figure 4. Execution time of 30 Lanczos iterations on a 933MHz Pentium III Beowulf cluster (solid line with symbols). The dashed lines illustrate ideal scaling. The nearest neighbor CPU communication limits the 0.25, 0.5, 1, and 2 million atom simulations to 32, 40, 51, and 63 CPUs, repectively. (b) Efficiency as defined as the ratio of actual speed-up to ideal speed-up.
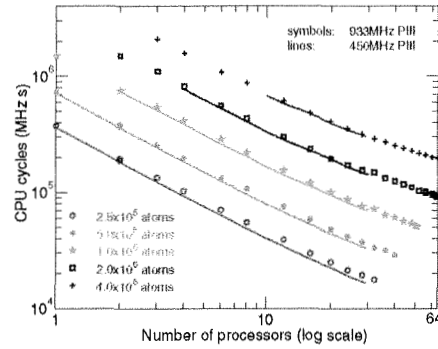

Figure 5. Comparison of a 450 MHz Beowulf with 512 MB of memory per node (solid) with a 933 MHz Beowulf with 1 GB of memory per node (symbols) in terms of CPU cycles.

The two largest problems require enough memory that they require more than one processor to avoid swapping. The efficiency, the ratio of speedup to ideal speedup, as a function of number of processors decreases with increasing network size. This decrease corresponds to a fraction of unparallelized code of approximately 1.6%. However, Figure 4b also shows that for the regime of interest the efficiency is independent of problem size. This result suggests that while communication bandwidth is not a limiting factor for this problem, there may be some inherent load imbalance in the computation. This issue is under continued investigation.

Figure 5 compares a 933 MHz Pentium III dual CPU Beowulf to a 450MHz single CPU Pentium III Beowulf. If execution times are scaled by CPU clock frequency, the performances of the two systems are indistinguishable. This result reinforces our finding that for the problem sizes of interest, our solver is limited by CPU speed rather than by communication bandwidth.

NEMO 3-D has the option of reusing the Hamiltonian for more than one iteration or computing the matrix vector multiplications on the fly without explicitly storing
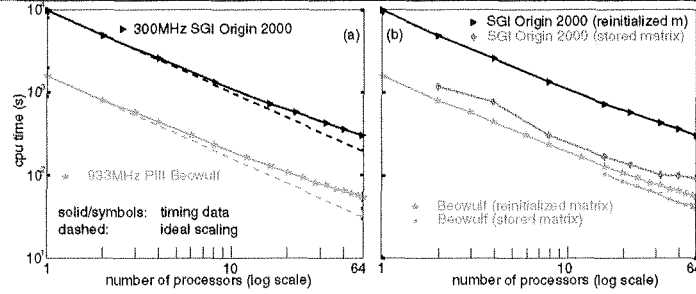
Figure 6. Comparison of CPU times on a Beowulf cluster with an SGI Origin 2000 with (a) ideal cpu time and (b) stored vs. recomputed Hamiltonian.

the Hamiltonian. Clearly, storing the Hamiltonian for later re-use is preferable, but is not possible for sufficiently large problems. Figure 6 compares the performance of the 933MHz cluster with a 128 CPU 300 MHz R12k SGI Origin 2000 with 512 MB of memory per processor. As before, dashed curves indicate ideal performance. Figure 6a simply shows that the SGI Origin 2000 scales similarly to the Beowulf cluster. Figure 6b, interestingly, shows that the benefit of storing the Hamiltonian is much greater for the Origin 2000 than it is for the Beowulf cluster. Indeed, while execution time on the Beowulf system is reduced only by a factor of 1.3, it is reduced roughly by a factor of 4 on the SGI Origin 2000. The reason for this discrepancy is not completely clear, but is likely attributable to a difference is cache size and memory access speed.

5. Simulation of Alloyed Quantum Dots

Since we represent each individual atom in the quantum dot system explicitly we can demonstrate this capability by simulating an $In_{0.6}Ga_{0.4}As$ alloyed quantum dot system[19] in a GaAs matrix. The dome shaped quantum dot has a diameter of 30nm and a height of 5nm. A 5nm GaAs buffer surrounded the quantum dot in the simulation. Since the In and Ga ions inside the alloyed dot are randomly distributed, different alloy configurations exist and optical transition energies from one dot to the next may vary, even if the size of the dot is assumed to be fixed. We therefore try to answer the question: What is the minimal optical linewidth that can be expected for such an alloyed dot neglecting any experimental size variations? A side view of such an alloyed quantum dot which is half the size of the system considered here without the surrounding GaAs is shown in Fig. 7a.

In simulations of 490 random alloy configurations[17] we have obtained the single particle electron and hole ground states energies and the optical transition energy from their difference. The mechanical strain is minimized using a valence force field method[20, 21] which considers contributions to the total strain energy due to bond length changes as well as bond angle modification. The mechanical strain field is recomputed for each alloy configuration. For these particular simulations we used the sp3s* basis set where the matrix elements scale with respect to the equilibrium position with the ideal[14] exponent 2.
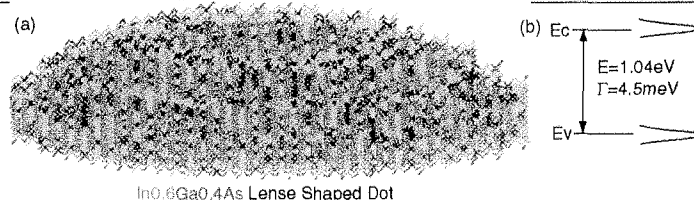
Figure 7. (a) Lens shaped $In_{0.6}Ga_{0.4}As$ quantum dot visualized the molecular viewer Rasmol. The quantum dot size is scaled down to include less than 10,000 atoms for visualization purposes. (b) Lowest conduction and highest valence quantized state that are inhomogeneously broadened by alloy disorder.

The simulation of 50 different alloyed dots shows a mean optical transition energy of 1.04eV and a standard deviation, or associated linewidth of 4.5meV compared to experimentally reported[19] transition energy of 1.09eV and a linewidth of 34.6meV. The experimental data does of course include quantum dot size variations as well. We plan to simulate larger samples that do include quantum dot size variations in the future. The major result of this simulation is the observation that there will be a significant optical linewidth variation due to alloy disorder alone, even if all the quantum dots were perfectly identical in size.

Acknowledgements

References

1. Data taken from The National Technology Roadmap for Semiconductors, Semiconductor Industry Association, San Jose, California, USA, 1998. Classifications such as CMOS Devices with Quantum effects and Quantum Devices taken from private communication with Alan Seabaugh, Texas Instruments (1997)
2. R. Chris Bowen, best case estimates for number of electrons under an SRAM cell, personal communication 1998, Texas Instruments.
3. H.C. Liu et al., Appl. Phys. Lett. 78, 79 (2001).
4. A.O. Orlov et al., Applied Physics Letters 78, 1625 (2001).
5. A. Canning, L.W. Wang, A. Williamson, and A. Zunger, Journal of Computational Physics 160, 29 (2000).
6. Physics-based device simulation tools have typically only been used to improve individual device performance after careful calibration of the simulation parameters.
7. R. Chris Bowen et al., J. Appl. Phys 81, 3207 (1997).
8. G. Klimeck et al., in the 1997 55th Annual Device Research Conference Digest, (IEEE, NJ, 1997), p. 92.
9. V. Ryzhii et al., Japanese Journal of Applied Physics, Part 2 Letters 39, L1283 (2000).
10. Seungwon Lee et al., Physical Review B in print (2001).
11. P. Vogl, H. P. Hjalmarson, and J. D. Dow, J. Phys. Chem. Solids 44, 365 (1983).

12. Timothy B. Boykin, Gerhard Klimeck, R. Chris Bowen, and Roger Lake, Phys. Rev. B 56, 4102 (1997).
13. Timothy B. Boykin, Phys. Rev. B 56, 9613 (1997).
14. J.M. Jancu, R. Scholz, F. Beltram, and F. Bassani, Phys. Rev. B 57, 6493 (1998).
15. Gerhard Klimeck *et al.*, Superl. and Microstr. 27, 77 (2000).
16. Gerhard Klimeck, R. Chris Bowen, Timothy B. Boykin, and Thomas A. Cwik, Superl. and Microstr. 27, 519 (2000).
17. G. Klimeck, F. Oyafuso, R. Chris Bowen, and Timothy B. Boykin, accepted for publication in Journal of Computational Electronics (2001).
18. J.C. Slater and G.F. Koster, Physical Review 94, 1498 (1954).
19. R. Leon *et al.*, Phys. Rev. B 58, R4262 (1998).
20. P. Keating, Phys. Rev. 145, 637 (1966).
21. C. Pryor *et al.*, J. of Appl. Phys. 83, 2548 (1998).